Biostatistics (2004), **5**, 1, *pp.* 129–143
Printed in Great Britain

Tests for multivariate recurrent events in the presence of a terminal event

BINGSHU ERIC CHEN*, RICHARD J. COOK
Department of Statistics and Actuarial Science, University of Waterloo, 200 University Avenue West, Waterloo, Ontario, Canada N2L 3G1
cheneric@mail.nih.gov

SUMMARY

In studies involving diseases associated with high rates of mortality, trials are frequently conducted to evaluate the effects of therapeutic interventions on recurrent event processes terminated by death. In this setting, cumulative mean functions form a natural basis for inference for questions of a health economic nature, and Ghosh and Lin (2000) recently proposed a relevant class of test statistics. Trials of patients with cancer metastatic to bone, however, involve multiple types of skeletal complications, each of which may be repeatedly experienced by patients over their lifetime. Traditionally the distinction between the various types of events is ignored and univariate analyses are conducted based on a composite recurrent event. However, when the events have different impacts on patients' quality of life, or when they incur different costs, it can be important to gain insight into the relative frequency of the specific types of events and treatment effects thereon. This may be achieved by conducting separate marginal analyses with each analysis focusing on one type of recurrent event. Global inferences regarding treatment benefit can then be achieved by carrying out multiplicity adjusted marginal tests, more formal multiple testing procedures, or by constructing global test statistics. We describe methods for testing for differences in mean functions between treatment groups which accommodate the fact that each particular event process is ultimately terminated by death. The methods are illustrated by application to a motivating study designed to examine the effect of bisphosphonate therapy on the incidence of skeletal complications among patients with breast cancer metastatic to bone. We find that there is a consistent trend towards a reduction in the cumulative mean for all four types of skeletal complications with bisphosphonate therapy; there is a significant reduction in the need for radiation therapy for the treatment of bone. The global test suggests that bisphosphonate therapy significantly reduces the overall number of skeletal complications.

Keywords: Marginal methods; Multivariate response; Recurrent event; Robust inference; Terminal event.

1. INTRODUCTION

In studies involving diseases associated with high mortality rates, trials are frequently conducted to evaluate the effects of therapeutic interventions on response processes terminated by death. Examples include studies of medical costs (Lin *et al.*, 1997; Bang and Tsiatis, 2000), quality of life (Zhao and Tsiatis, 1997, 1999) and recurrent events (Cook and Lawless, 1997; Li and Lagakos, 1997). Interest typically lies in cumulative aspects of such response processes, such as the cumulative lifetime costs, quality adjusted lifetime, or the cumulative lifetime number of events. Analyses dealing with such

*To whom correspondence should be addressed.

questions must address the dependent censoring of the cumulative response which results from the fact that survival times are typically right-censored for some individuals (Strawderman, 2000). Suitable techniques are frequently formulated in terms of ‘inverse probability of censoring weighted’ analyses, although alternative approaches can also be taken (e.g. Strawderman, 2000).

For problems in which a single type of recurrent event is of interest, Cook and Lawless (1997) proposed the use of cumulative mean functions which reflect the marginal cumulative mean number of events experienced per patient over time, accounting for the fact that the recurrent event process is terminated by death. Such mean functions form a natural basis for inference for questions of a health economic nature or for other settings where interest lies in comparing overall disease burden at the population level. If interest lies in testing for differences in cumulative mean functions between groups, a class of test statistics recently developed by Ghosh and Lin (2000) may be used.

Trials of patients with cancer metastatic to bone, however, involve multiple types of skeletal complications which may be repeatedly experienced by patients over the course of follow-up (Theriault *et al.*, 1999). Traditionally the distinction between the various types of events is ignored and univariate analyses are conducted based on a composite recurrent event. However, when the events have a different impact on quality of life, or when they incur different costs, it can be important to gain insight into the relative frequency of the various types of events and treatment effects thereon. This may be achieved by conducting separate marginal analyses with each analysis focusing on one type of recurrent event. Global inferences regarding treatment effect can then be conducted by carrying out multiplicity adjusted marginal tests, using more formal multiple testing procedures, or by constructing global test statistics. We describe methods for testing for differences between treatment groups with respect to multiple cumulative mean functions in the presence of a common terminal event (i.e. death). These methods will be shown to be valid in settings where there is a dependence between the recurrent event rate and survival time; naive tests based on rate functions (e.g. Cook *et al.*, 1996) are invalid in such settings.

The remainder of the paper is organized as follows. In Section 2 we briefly discuss a motivating trial which was designed to assess the effect of bisphosphonate therapy on the occurrence of various skeletal complications in patients with breast cancer metastatic to bone (Theriault *et al.*, 1999). We develop the methods in Section 3 and report on the results of simulation studies in Section 4. The methods are applied to the data from Theriault *et al.* (1999) in Section 5. Concluding remarks and topics for future research are given in Section 6. Note that while our focus is on problems with multiple types of recurrent events, the methods we describe are easily adapted to deal with multi-type quality of life data which arises by separately considering different domains of quality of life, or multi-type cost data which would arise if it was of interest to separately examine personnel costs, drug costs, etc.

2. BISPHOSPHONATES FOR THE TREATMENT OF BONE METASTASES

Theriault *et al.* (1999) report on a multicenter randomized trial designed to investigate the effect of a bisphosphonate, pamidronate, on the development of skeletal complications in breast cancer patients with bone metastases. Patients were accrued from 85 study sites in the United States, Canada, Australia and New Zealand. Patients with stage IV breast cancer receiving cytotoxic chemotherapy with at least two predominantly lytic bone lesions at least one centimeter in diameter were randomized within strata defined by ECOG status. A total of 371 women were enrolled in the study with 182 randomized to receive 90 mg of pamidronate every four weeks and 189 randomized to receive dextrose infusions at the same time points. After completion of the planned one year of follow-up, the follow-up was extended for an additional year to assess long-term effects and survival. At monthly visits patients were assessed and the occurrence of skeletal complications was recorded. Skeletal complications include nonvertebral and vertebral fractures, the need for surgery to treat or prevent fractures, and the need for radiation for the

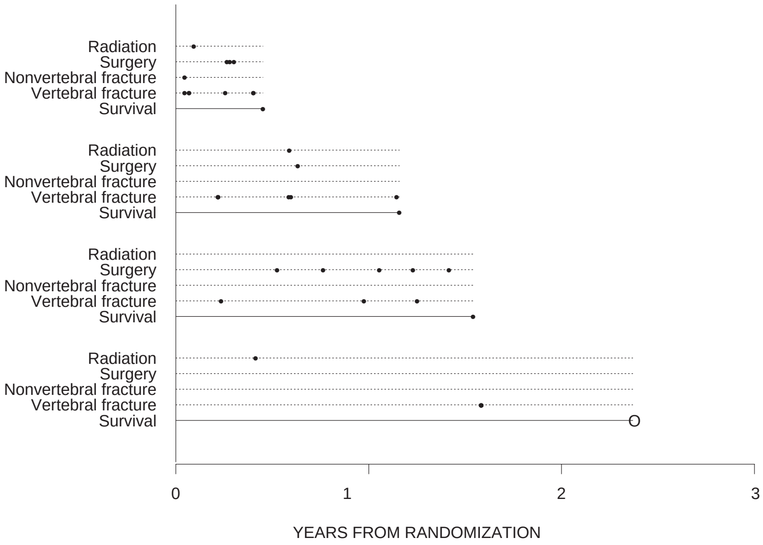

Fig. 1. Sample profile features for various skeletal complications and survival (open circles represent right censored survival times).

treatment of bone pain. Each patient was followed until death, the last date of contact, or loss to follow-up. Figure 1 displays the timing and number of skeletal related events for four patients from Theriault *et al.* (1999). This figure illustrates the fact that each patient may experience more than one of each type of event prior to death, and that some deaths are right-censored. The focus here is (i) to estimate the cumulative mean number of events of each type over time in the control and treatment arms, (ii) to examine the effect of bisphosphonate therapy on the cumulative mean number of events of each type, and (iii) to carry out global assessments of the effect of bisphosphonate therapy.

3. Multi-type recurrent events with dependent termination

3.1 *Methods for univariate recurrent events*

Suppose there are a total of m subjects accrued into a study and each subject is at risk for a particular type of recurrent event. Let $(0, \tau]$ represent the period of observation and let $N_i^*(u)$ be a right-continuous integer function representing the number of events experienced by subject i over the interval $(0, u]$, $i = 1, 2, \ldots, m$, $0 < u \leqslant \tau$. During the observation period $(0, \tau]$, some subjects may experience an event which terminates the recurrent event processes (e.g. death), but they may also withdraw from the study according to some random censoring mechanism which is independent of the recurrent event and terminal event processes. For $i = 1, \ldots, m$, let T_i be the time of the terminating event, C_i the censoring time, $X_i = \min(T_i, C_i)$, $\pi(t) = \Pr(X_i \geqslant t)$, and $\delta_i = I(X_i = T_i)$, where $I(\cdot)$ is the indicator function. We let $N_i(t) = N_i^*\{\min(t, T_i)\}$ denote the number of recurrent events observed over $(0, t]$ in the presence of death. The data contributed by each subject then take the form $(\{N_i(u), 0 < u \leqslant X_i\}, X_i, \delta_i)$, $i = 1, \ldots, m$. Let $Y_i(t) = I(X_i > t)$ be the at risk indicator function which is one when subject i is under observation and at risk for the recurrent events at time t, and is zero otherwise. We suppose initially that we have a single sample of subjects.

A marginal approach for the analysis of recurrent events in the presence of a dependent terminal event was proposed in Cook and Lawless (1997). In this approach the distribution of the failure time is modeled marginally, but the rate function for the recurrent events is specified conditionally on the subject not having experienced the terminal event before some time point t. Thus we may write the rate of events for subject i at time t, conditional on them not having experienced the terminal event up to time t, as

$$\mathrm{d}R(t) = E\{\mathrm{d}N_i(t)|T_i \geqslant t\}, \quad 0 \geqslant t. \tag{3.1}$$

Then, the expected number of events over the interval $(0, t]$, conditional on the terminal event not occurring before t (i.e. $R(t) = E\{N_i(t)|T_i \geqslant t\}$), can be estimated via a Nelson–Aalen type estimate as

$$\hat{R}(t) = \frac{\sum_{i=1}^{m} Y_i(t)N_i(t)}{\sum_{i=1}^{m} Y_i(t)}. \tag{3.2}$$

The marginal expectation for the number of recurrent events over the interval $(0, t]$, accounting for possible termination of the event process, is given by

$$\mu(t) = \mathrm{E}\{N(t)\} = \int_0^t P(T \geqslant u)\ E\{\mathrm{d}N(u)|T \geqslant u\}\mathrm{d}u = \int_0^t S(u)\mathrm{d}R(u), \tag{3.3}$$

where $S(u)$ is survival function of terminal event times. An estimate of $\mu(t)$ is given by

$$\hat{\mu}(t) = \int_0^t \hat{S}(u)\mathrm{d}\hat{R}(u), \tag{3.4}$$

where $\mathrm{d}\hat{R}(u) = \sum_{i=1}^{m} Y_i(u)\,\mathrm{d}N_i(u)/\sum_{i=1}^{m} Y_i(u)$ and $\hat{S}(u)$ is the Kaplan–Meier estimate of the survival function (Cook and Lawless, 1997).

To make inferences about differences between two groups (say a treatment and a control group), tests may be constructed based on marginal cumulative mean functions. Let z_i be a binary covariate such that $z_i = 1$ for subjects in the treated group and $z_i = 0$ for subjects in the control group, and let $\mu_1(t)$ and $\mu_0(t)$ denote the marginal mean functions for the treatment and control groups respectively, where $\mu_\ell(t) = \int_0^t S_\ell(u)\mathrm{d}R_\ell(u),\ \ell = 0, 1$. Note that the survival distributions for the control and treatment groups need not be the same. It is therefore important to interpret $\mu_\ell(t)$ as the marginal cumulative mean number of events in group ℓ, adjusted by the fact that the recurrent event process is terminated by death. The mean function may be lower in one group due to a lower conditional rate function $\mathrm{d}R(u)$ or a higher mortality rate. Insight into which of these reasons explain any apparent differences can be gained from simultaneous consideration of Kaplan-Meier plots of the survival function $\hat{S}(t)$, the naive estimate of the cumulative mean function $\hat{R}(t)$, and the estimate of the marginal cumulative mean function $\hat{\mu}(t)$. The latter is a population attribute which is of interest in settings where the objective is to assess differences in the overall burden of disease such as when costs are associated with events.

Suppose the null hypothesis is that there is no difference in the marginal cumulative mean functions between the treatment and control groups,

$$H_0:\ \mu_0(t) = \mu_1(t),\ \text{for } 0 < t \leqslant \tau.$$

Ghosh and Lin (2000) consider tests for this hypothesis based on a generalized log-rank statistic of the form

$$\hat{Q} = \left(\frac{m_0 m_1}{m}\right)^{\frac{1}{2}} \int_0^\tau \hat{W}_m(t)\mathrm{d}\{\hat{\mu}_1(t) - \hat{\mu}_0(t)\}, \tag{3.5}$$

where $\hat{W}_m(t)$ is a weight function based on observed data. Our aim is to extend this methodology to deal with settings in which there are multiple types of recurrent events.

3.2 *Tests for multi-type recurrent events*

If there are K types of recurrent events, let $N_{ki}(u)$ record the number of recurrent events of type k occurring over the interval $(0, u]$, $k = 1, \ldots, K$. One strategy for dealing with this multivariate recurrent event is to construct a composite event which is said to have occurred if any one of the particular types of events occurs. We can let

$$N_{\cdot i}(u) = N_{1i}(u) + N_{2i}(u) + \cdots + N_{Ki}(u).$$

Such an approach is somewhat appealing in the sense that the multivariate process is reduced to a univariate process and hence methodology discussed by Ghosh and Lin (2000) can be applied to the recurrent event process $N_{\cdot i}(u)$ directly. We denote the Ghosh and Lin (2000) test statistic based on this composite event by $\hat{Q}_{\cdot}$ and let $\bar{Q}_{\cdot} = \hat{Q}_{\cdot}/\sqrt{\widehat{\text{var}}(\hat{Q}_{\cdot})}$ represent the standardized form.

As discussed earlier, however, it may be desirable to study the relative frequency of different types of events and to understand how a treatment affects their occurrence. In this case one may wish to form marginal models in which each type of recurrent outcome is analyzed separately. The evidence of treatment effects may then be interpreted marginally and in addition, one may draw simultaneous inferences about the overall treatment effect. For example, multiplicity adjusted p-values may be constructed to test the null hypothesis of no overall effect of treatment based on the marginal p-values observed. Alternatively, the global statistics may be constructed to test for evidence against the null hypothesis of no treatment effect over all event types. These are the approaches taken here, but remarks are subsequently made regarding the use of composite events in the simulation studies.

Consider a single sample of m subjects. Let $\mathrm{d}R_k(t) = E\{\mathrm{d}N_{ki}(t)|T_i \geqslant t\}$ and $\mu_k(t) = E\{N_{ki}(t)\}$ denote the conditional rate function and cumulative mean function for events of type k. These can be estimated from (3.2) and (3.4) by replacing $\mathrm{d}N_i(u)$ with $\mathrm{d}N_{ki}(u)$, for $k = 1, \ldots, K$, and $i = 1, \ldots, m$. Let $\boldsymbol{\mu}(t) = \{\mu_1(t), \mu_2(t), \ldots, \mu_K(t)\}'$ and $\hat{\boldsymbol{\mu}}(t) = \{\hat{\mu}_1(t), \hat{\mu}_2(t), \ldots, \hat{\mu}_K(t)\}'$. For each type of recurrent event, Ghosh and Lin (2000) show that the difference between $\hat{\mu}_k(t)$ and $\mu_k(t)$ can be approximated by the sum of independent and identically distributed random variables. Specifically,

$$\sqrt{m}\{\hat{\mu}_k(t) - \mu_k(t)\} = \frac{1}{\sqrt{m}} \sum_{i=1}^{m} \Psi_{ki}(t) + o_p(1),$$

where the expression for $\Psi_{ki}(t)$ is given in expression (A.1) of the Appendix. For fixed t, $\Psi_{ki}(t)$ are independent and identically distributed mean zero random variables and we show in the Appendix that $\sqrt{m}\{\hat{\boldsymbol{\mu}}(t) - \boldsymbol{\mu}(t)\}$ follows a mean zero asymptotic multivariate normal distribution. The asymptotic covariance between the jth component and kth component is consistently estimated by

$$\hat{\xi}_{jk} = \frac{1}{m} \sum_{i=1}^{m} \{\hat{\Psi}_{ji}(t)\hat{\Psi}_{ki}(t)\};\ 1 \leqslant j, k \leqslant K,$$

where $\hat{\Psi}_{ki}(t)$ is obtained by replacing all the unknown quantities of $\Psi_{ki}(t)$ with their corresponding empirical estimates. Let $m_\ell = m_\ell(0)$ denote the number of subjects initially in group ℓ, $\ell = 0$ or 1, $m = m_0 + m_1$, and $\hat{\mu}_{k\ell}(t)$ is the estimate of the mean function for event k and group ℓ from (3.4) by replacing $\mathrm{d}N_i(u)$ with $\mathrm{d}N_{ki}(u)$. For the two sample problem a marginal test statistic $\hat{Q}_k$ can be obtained from (3.5) by replacing $\mathrm{d}\hat{\mu}_\ell(t)$ with $\mathrm{d}\hat{\mu}_{k\ell}(t)$ giving

$$\hat{Q}_k = \left(\frac{m_0 m_1}{m}\right)^{\frac{1}{2}} \int_0^\tau \hat{W}_m(t)\mathrm{d}\{\hat{\mu}_{k1}(t) - \hat{\mu}_{k0}(t)\}, \tag{3.6}$$

where the weight function is given by

$$\hat{W}_m(t) = \frac{m_0(t)m_1(t)}{m(t)}\frac{m}{m_0 m_1},$$

with $m_\ell(t) = \sum_{i=1}^{m} I(z_i = \ell)Y_i(t)$ for $\ell = 0$ and 1, $m(t) = m_0(t) + m_1(t)$, and where z_i is a binary covariate. If $m_0/m \to \rho_0$, $m_1/m \to \rho_1$ as $m \to \infty$, then

$$\lim_{m\to\infty} \hat{W}_m(t) = \frac{\pi_0(t)\pi_1(t)}{\pi(t)} = W(t),$$

where $\pi_\ell(t) = \Pr\{Y_i(t) = 1, z_i = \ell\}$ for subject i in group ℓ, and $\pi(t) = \pi_0(t) + \pi_1(t)$ (Ghosh and Lin, 2000).

Suppose that it is of interest to test the overall null hypothesis of no treatment effect on any of the recurrent event types. If

$$H_{k0}: \ \mu_{k0}(t) = \mu_{k1}(t),\ k = 1, 2, \ldots, K;\ 0 < t \leqslant \tau,$$

are the separate null hypotheses, then $H_0 = \bigcap_{k=1}^{K} H_{k0}$ is the overall null hypothesis. Let $\hat{\boldsymbol{Q}} = (\hat{Q}_1, \ldots, \hat{Q}_K)'$. Since $\hat{Q}_k$ is a function of type k events and the events are correlated, the marginal test statistics $\hat{Q}_k$, $k = 1, 2, \ldots, K$ are correlated. Thus, under the null hypotheses that treatment does not affect any type of recurrent event, $\hat{\boldsymbol{Q}}$ asymptotically follows a multivariate normal distribution with mean zero and the covariance matrix Σ, where the (j, k) entry of Σ is consistently estimated by

$$\widehat{\text{cov}}(\hat{Q}_j, \hat{Q}_k) = \frac{1}{m}\sum_{\ell=0}^{1}\frac{m_{1-\ell}}{m_\ell}\left[\sum_{i=1}^{m}\left\{\int_0^\tau \hat{W}_m(t)\mathrm{d}\hat{\Psi}_{ji}(t)\int_0^\tau \hat{W}_m(t)\mathrm{d}\hat{\Psi}_{ki}(t)\right\} I(z_i = \ell)\right], \tag{3.7}$$

for $1 \leqslant j, k \leqslant K$. We let $\hat{\Sigma}$ denote the estimator of Σ, and $\bar{Q}_k = \hat{Q}_k/\sqrt{\widehat{\text{var}}(\hat{Q}_k)}$ denote the standardized form of the test statistic $\hat{Q}_k$ (see Appendix). The correlation matrix of $(\bar{Q}_1, \bar{Q}_2, \ldots, \bar{Q}_K)$ is consistently estimated by $\hat{\Gamma} = \{\text{diag}(\hat{\Sigma})\}^{-\frac{1}{2}}\hat{\Sigma}\{\text{diag}(\hat{\Sigma})\}^{-\frac{1}{2}}$. Given these distributional results one can construct a test statistic given by $\hat{\boldsymbol{Q}}'\hat{\Sigma}^{-1}\hat{\boldsymbol{Q}}$, which asymptotically follows a χ^2_K distribution under the null hypothesis. For the purpose of making treatment comparisons, however, more directed tests are desirable as discussed in what follows (O'Brien, 1984).

Now that we have the marginal statistics $\hat{Q}_1, \hat{Q}_2, \ldots, \hat{Q}_K$ and the covariance matrix among these statistics, a global test statistic regarding treatment effects can be obtained as a special case of Rao (1973, Section 1.f), on the optimal linear combination of several estimators. Specifically we take

$$\hat{Q}_w = c_1\bar{Q}_1 + c_2\bar{Q}_2 + \cdots + c_K\bar{Q}_K, \tag{3.8}$$

where $\boldsymbol{c} = (c_1, c_2, \ldots, c_K)$ satisfies $c_1 + c_2 + \cdots + c_K = 1$. The term 'optimal' means here that such a linear combination will minimize the asymptotic variance among all linear transformations subject to the restriction $c_1 + c_2 + \cdots + c_K = 1$. Here the weight function given by $\boldsymbol{c} = (\boldsymbol{J}'\hat{\Gamma}^{-1}\boldsymbol{J})^{-1}\hat{\Gamma}^{-1}\boldsymbol{J}$, where $\boldsymbol{J} = (1, 1, \ldots, 1)'$ satisfies this optimality criterion. A simple calculation shows that under the null hypothesis, $\hat{Q}_w$ asymptotically follows a normal distribution with mean zero and asymptotic variance estimated by $\widehat{\text{var}}(\hat{Q}_w) = (\boldsymbol{J}'\hat{\Gamma}^{-1}\boldsymbol{J})^{-1}$. We let $\bar{Q}_w = \hat{Q}_w/\sqrt{\widehat{\text{var}}(\hat{Q}_w)}$ denote the standardized form of this weighted global statistic. This global statistic is similar in spirit to that proposed by Wei *et al.* (1989) for multivariate survival data. One appeal to the construction of global test statistics is the fact that weights may be chosen to reflect the relative severity or costs of the events, rather than solely on the basis of precision considerations.

An alternative way of testing the overall null hypothesis $H_0 : \bigcap_{k=1}^{K} H_{k0}$ is based on marginal test statistics. If the overall type I error rate is α, and interest lies in making separate inferences regarding each event type, then Bonferroni adjustments can be used. This would involve testing each null hypothesis H_{k0}, $k = 1, \dots, K$, at the nominal level α/K. Bonferroni adjustments are well known to be conservative with respect to the overall type I error, and if interest only lies in testing a global hypothesis, a less conservative improvement proposed by Simes (1986) can lead to more powerful tests. In Simes' procedure the marginal p-values are ordered from smallest to largest and denoted $p_{(1)} < \cdots < p_{(K)}$. If $p_{(k)} < k\alpha/K$ for any $k = 1, \dots, K$ then the null hypothesis H_0 is rejected and overall type I error rate is preserved. More sophisticated less conservative methods (e.g. Armitage and Parmar, 1986, James, 1991) can be employed based on the correlation between the test statistics, but with small to modest correlations Simes (1986) strikes a good balance between simplicity and performance.

To compare the proposed methods with other nonparametric testing procedures, we consider the use of a naive test based on the pseudo score statistic in Cook *et al.* (1996) given by

$$\hat{Q}_N = \int_0^{\tau} W(t)\{\mathrm{d}\hat{\Lambda}_1(t) - \mathrm{d}\hat{\Lambda}_2(t)\},$$

where $\hat{\Lambda}_j(t) = \sum_{i=1}^{m} \int_0^t Y_i(u)\mathrm{d}N_i(u)/Y_{.}(u)$, j = 1, 2, are naive Nelson–Aalen estimates for the cumulative mean function of the recurrent events. This naive standardized test statistic is denoted by $\bar{Q}_N$. All statistics that we discuss here are based on one degree of freedom.

4. Simulation studies

Here we assess the finite sample performance of various approaches to the analysis by specifying particular models for the recurrent and terminal events. For simplicity we consider $K = 2$ correlated recurrent event processes which are terminated by a common terminal event, representing death for example. Let $(u_i, v_i)'$ denote a bivariate normal random variable with $E(u_i) = E(v_i) = 0$, $\mathrm{var}(u_i) = \sigma_1^2$, $\mathrm{var}(v_i) = \sigma_2^2$, and $\mathrm{corr}(u_i, v_i) = \rho$. Conditional on u_i, let the terminal event time T_i be taken to follow a log normal distribution, so

$$\{\log(T_i)|u_i, v_i\} \sim N(\mu_i, \sigma_T^2),$$

where $\mu_i = \mu_0 + u_i$ is the mean, μ_0 is the intercept. This implies that unconditionally $\log(T_i)$ follows a normal distribution with mean μ_0 and variance $(\sigma_T^2 + \sigma_1 v^2)$. An independent right censoring time C_i is taken to follow an exponential distribution with mean ψ where ψ can be chosen to give varying degrees of censoring for the terminal event time. Conditional on v_i, type k recurrent events are taken to follow a Poisson process with rate function

$$\lambda_{ki}(s|v_i, z_i) = \lambda_{k0}(s)\exp(z_i'\beta_k + v_i), \quad 0 \leqslant s \leqslant t_i, \; i = 1, \dots, m, \quad k = 1, 2,$$

where $\lambda_{k0}(s)$ is a baseline rate function, $z_i = 1$ if subject i is in the treatment group and $z_i = 0$ otherwise. Here β_k reflects the effect of treatment on events of type k, $k = 1, 2$ and v_i models residual heterogeneity in the rate of events among subjects within the same treatment group as well as the association between type 1 and 2 recurrent events. The association between the recurrent events and the time of the terminal event is modeled by the correlation $\rho = \mathrm{corr}(u_i, v_i)$.

Data for $m = 200$ subjects were simulated with 100 subjects in the treatment and control groups respectively observed over the interval $(0, \tau]$ where without loss of generality we take $\tau = 1$. For the survival time T, we let $\mu_0 = 0.5$ and $\sigma_T^2 = 0.1$. We adopt time-homogeneous baseline event rates of $\lambda_k = 4$ or 8, $k = 1, 2$ to represent moderately frequent and more frequent recurrent events. We let $\beta_k = 0$ when we want no treatment effect on the rate of events for the kth process, and let $\beta_k = \log(0.75)$ to denote

Table 1. *Empirical type I error rates (percent) for naive tests* ($\bar{Q}_N$), *univariate analyses* ($\bar{Q}_1$ *and* $\bar{Q}_2$), *composite endpoint analyses* ($\bar{Q}_.$), *global tests* $\bar{Q}_w$), *Bonferroni adjusted tests* (BONF) *and Simes procedure* (SIMES) *under the null hypothesis* $H_0 : \mu_{k1}(t) = \mu_{k0}(t), k = 1, 2$.[†] (*Based on 2000 simulations with sample size* $m = 200$)

λ_1	λ_2	CEN%	ρ	$\bar{Q}_N$	$\bar{Q}_1$	$\bar{Q}_2$	$\bar{Q}_.$	$\bar{Q}_w$	BONF	SIMES
4	4	25	0.25	15.9	4.5	5.1	4.1	4.2	4.9	4.9
8	4	25	0.25	18.5	5.4	5.1	5.4	5.2	5.1	5.2
8	8	25	0.25	17.2	6.3	5.3	5.8	5.9	5.7	5.8
4	4	50	0.25	18.1	5.1	5.4	5.1	5.2	5.9	5.2
8	4	50	0.25	17.5	5.5	5.1	5.3	5.2	5.3	5.4
8	8	50	0.25	20.6	5.1	4.8	5.5	5.8	4.6	4.7
4	4	25	0.75	32.9	5.4	5.2	5.0	5.5	4.8	4.9
8	4	25	0.75	31.1	5.2	5.0	4.9	5.0	4.4	4.7
8	8	25	0.75	27.9	5.5	5.1	5.7	5.6	4.7	5.1
4	4	50	0.75	36.6	5.1	5.4	5.4	5.4	4.5	4.8
8	4	50	0.75	33.3	4.9	4.6	4.9	4.2	4.2	4.1
8	8	50	0.75	35.7	5.6	5.7	5.9	5.8	4.7	5.0

[†] Note: λ_k indicates the marginal expected number of type k event, CEN% indicates the percentage of subjects censored at the end of the study, and ρ is the correlation coefficient between the random effects.

a treatment effect in which the event rate in the treatment group is three-quarters that of the control group, $k = 1, 2$. We restrict consideration here to scenarios with shared random effects between the recurrent event processes which leads to a positive association between recurrent events, and set $\sigma_1 = 1.0$. The correlation between random effect u_i and v_i is set at -0.25 and -0.75, to correspond to lower rates of recurrent events for subjects with lower hazards for the terminal event. Finally, we change the value of the parameter ψ in the censoring distribution to obtain a 25% and 50% censored data for the time of the terminal event. The simulated observable data takes the form $(\{N_{ki}(u), 0 \leqslant u \leqslant X_i, k = 1, 2\}, X_i, \delta_i, z_i)$, $i = 1, 2, \ldots, 200$.

Let $\bar{Q}_1$, $\bar{Q}_2$ denote the standardized univariate Ghosh and Lin (2000) test statistics based on (3.5) and $\bar{Q}_.$ denote the corresponding statistic based on the composite event analysis. The statistic $\bar{Q}_w$ represents the weighted global test statistic based on (3.8).

First we examine the empirical type I error rates for the tests based on $\bar{Q}_N$, $\bar{Q}_1$, $\bar{Q}_2$, $\bar{Q}_.$, $\bar{Q}_w$, univariate tests based on Bonferroni adjustments (BONF), and Simes' multiple testing procedures (SIMES), where $\bar{Q}_N$ is the naive statistic, $\bar{Q}_k, k = 1, 2$ are univariate test statistics based on the two event types, $\bar{Q}_.$ is the test statistic based on the composite recurrent event process defined by summing the counts for the individual processes, and the global test statistic $\bar{Q}_w$ is based on the formula in Section 3.2. To generate the data set, we let $\beta_1 = \beta_2 = 0$. For each combination of the baseline recurrent event rate and degree of censoring for the terminal event, 2000 data sets are generated. For each data set, each procedure for testing was carried out based on a two-sided test of the null hypothesis at the 5% significance level. The empirical event rate is computed as the proportion of data sets in which the null hypothesis was rejected by the corresponding test. From Table 1 we can see that all tests except for the naive test have empirical type I error rates which are compatible with the nominal 5% level. We conclude that for the purpose of constructing linear global tests, the expressions for the covariance estimates of the marginal test statistics appear reasonable for the finite sample settings considered here. Furthermore, note that the test ignoring the effect of termination is associated with an empirical type I error rate considerably greater than the

Table 2. *Empirical power (percent) under alternative hypothesis (with at least one or both of β_k equal to* $\log(3/4) = -0.287$) *for Bonferroni adjusted tests* (BONF), *Simes procedures* (SIMES), *composite endpoint analyses* ($\bar{Q}$), *and global tests* ($\bar{Q}_w$) † (*based on 2000 simulations with sample size* $m = 200$)

					$\rho = 0.25$				$\rho = 0.75$			
λ_1	λ_2	β_1	β_2	CEN%	BONF	SIMES	$\bar{Q}$	$\bar{Q}_w$	BONF	SIMES	$\bar{Q}_{\cdot}$	$\bar{Q}_w$
4	4	0.0	−0.287	25	81.4	81.5	43.3	55.2	64.8	64.9	28.4	39.7
4	4	0.0	−0.287	50	77.0	77.2	39.4	50.1	61.0	60.9	27.4	38.8
4	4	−0.287	−0.287	25	92.7	93.1	96.3	96.3	79.5	80.3	84.7	84.3
4	4	−0.287	−0.287	50	90.5	91.4	94.8	94.8	76.7	77.2	82.9	82.8
8	4	0.0	−0.287	25	81.3	81.4	24.9	82.6	66.6	66.7	15.5	81.9
8	4	0.0	−0.287	50	77.6	77.9	23.1	78.9	64.1	64.1	17.0	80.6
8	4	−0.287	0.0	25	93.7	93.9	73.4	16.8	79.3	72.3	51.4	4.3
8	4	−0.287	0.0	50	92.3	92.5	70.2	15.4	73.4	73.4	47.8	7.2
8	4	−0.287	−0.287	25	97.2	87.8	98.4	94.8	83.8	76.2	87.4	73.1
8	4	−0.287	−0.287	50	94.8	83.0	97.4	91.1	82.0	71.0	86.1	66.3
8	8	0.0	−0.287	25	94.2	94.3	52.7	69.6	77.4	77.6	32.5	55.8
8	8	0.0	−0.287	50	92.5	92.5	51.1	66.1	73.0	73.1	30.5	51.6
8	8	−0.287	−0.287	25	97.8	98.2	99.2	99.0	86.2	87.5	89.4	89.2
8	8	−0.287	−0.287	50	97.6	97.9	98.6	98.6	83.4	84.7	86.8	86.8

†Note: λ_k indicates the marginal expected number of type k event, $k = 1, 2$, and ρ is the correlation coefficient between the random effects

nominal 5% level. Thus, with the presence of terminal events, the variance estimate of the naive test statistic is conservative.

Table 2 reports the empirical power of tests based on Bonferroni adjustments (BONF), Simes' multiple testing procedure (SIMES), the test statistic based on a univariate analysis of a composite event ($\bar{Q}_{\cdot}$), and the optimally weighted test statistic ($\bar{Q}_w$), under a variety of settings. In broad terms, the adjusted marginal procedures give higher power than the analyses based on the composite event or global test statistics when treatment effects are on one of the two events. If there are equal baseline rates for the recurrent events and the treatment effect is the same for the processes, then analyses based on the composite event and the global test statistic have comparable empirical power. This is in fact to be expected, since the test statistic based on the composite event is equivalent to a weighted statistic with equal weights on all components. Moreover, for recurrent events with equal baseline rates, common treatment effects induce equal variation for each component test statistic, in turn giving equal weights for the two components in the global test statistics. In summary, when the baseline event rates and the treatment effects are the same across event types the composite event analysis and the analysis based on the global test statistic have comparable power. These conclusions stand for both $\rho = 0.25$ and $\rho = 0.75$.

Another setting arises when the baseline rate functions for the recurrent processes are the same but the treatment effects on the two processes are different. For example, it may be that there is no treatment effect for one type of event but the treatment may reduce the rate of the other type of event. The results in Table 2 show that the optimal global test will give much higher power to detect such treatment effects under these parameter configurations. This increased power arises because when the treatment has no effect on one of the processes in the composite event this process introduces additional 'noise' but no 'signal', leading to a loss of power. This loss can be avoided in the global test statistic since the process for which the treatment reduces the rate of events has fewer expected events; this means there is less variation associated with

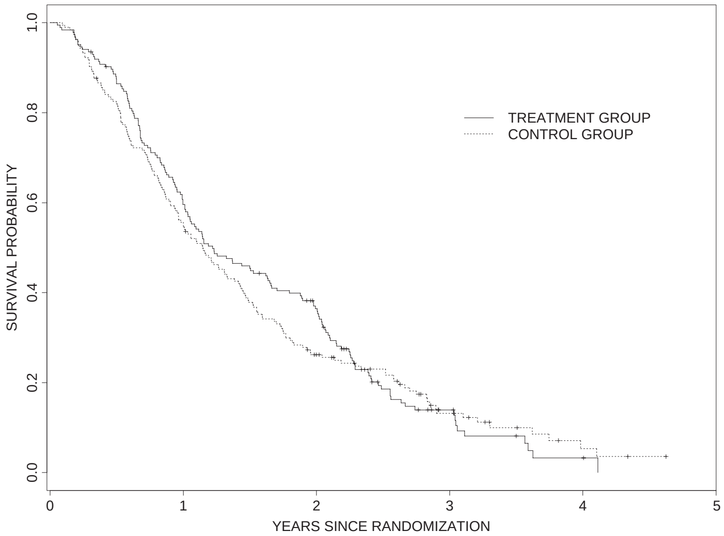

Fig. 2. Kaplan–Meier estimates of the survival functions for data from Theriault *et al.* (1999) (the (+) signs in the plot indicate right-censored survival times.)

this statistic (for a mixed Poisson process, the variance will increase when the mean increases). Hence the global test statistics will put more weight on the process with the treatment effect and less weight on the process without the treatment effect, thereby providing greater power. Of course one must carefully consider the results from such global tests since their interpretation is tied heavily to the weights used in forming the linear combination.

When the baseline rate functions are different, there are three ways in which the treatment may affect the event rates: (i) treatment effects may be equal for both processes, (ii) the treatment may only affect the rarer event, and (iii) the treatment effect may only reduce the rate of the more frequent event. In case (i) the analysis based on the composite event has higher power since the global test statistic will put more weight in the process with the lower expected number of events (because it has less variation), but the process with more frequent event has more power to detect the treatment effect, thus, the composite statistic has higher power. In case (ii) the global test statistic has higher power since it puts more weight on the process with the treatment effect. On the other hand, for case (iii), the global test statistic has more weight on the process without treatment effect, thus the composite test statistic has higher power.

This simulation study shows that the composite method and the global method are equivalent when the baseline rates and treatment effects are equal for each type of recurrent event. We prefer the global method when the baseline rates are equal and treatment effects are different or the treatment effect is on the process with less events. The composite method can be used if baseline event rates are different and treatments are equal or treatment effect is on the more frequent events.

5. APPLICATION TO THE BISPHOSPHONATE TRIAL

We now analyze the data from the study reported in Theriault *et al.* (1999). Here the recurrent events are monitored up to the time of loss to follow-up, death, or $\tau = 24$ months, whichever comes first. Figure 2 displays the Kaplan–Meier estimates of the survival functions for the control and treatment groups, revealing broadly similar distributions. This is not surprising since pamidronate is designed to improve patient quality of life by reducing the incidence of bone complications, rather than survival. The

Table 3. *Marginal analyses of the effect of bisphosphonates on non-vertebral fractures, vertebral fractures, need for radiation, and need for surgery for data from Theriault* et al. *(1999)*

		$E\{N_j(\tau)\}^\dagger$		Test			Optimal
		Pamidronate	Placebo	Statistics	Var	p-value	weights c_k
Non-vertebral	$\hat{Q}_1$	0.539	0.752	−0.109	0.126	0.388	0.232
Vertebral	$\hat{Q}_2$	0.805	1.004	−0.112	0.193	0.562	0.189
Radiation	$\hat{Q}_3$	0.573	1.059	−0.322	0.099	0.001	0.241
Surgery	$\hat{Q}_4$	0.076	0.145	−0.050	0.029	0.087	0.328
Global Test ‡	$\hat{Q}_w$			−0.144	0.069	0.037	
Composite	$\hat{Q}_.$	1.993	2.960	−0.593	0.318	0.062	
Composite	$\hat{Q}_N$	2.814	4.879	−1.399	0.162	<0.001	

$^\dagger E\{N_j(\tau)\}$ indicates the marginal expected number of recurrents for both the treatment and control groups.

‡Global test is based on weighted of the standardized version of $\hat{Q}_k$, $k = 1, 2, 3, 4$, with the weights given by the column c_k.

log-rank test gives $p = 0.635$ indicating insufficient evidence to reject the hypothesis of no treatment effect on survival. In the analyses that follow, however, we accommodate possible differences in the survivor functions by estimating the marginal mean functions based on (3.4) separately for each group.

Table 3 provides the summary $\hat{Q}$ statistics for the marginal analyses of each type of recurrent event as well as based on the Ghosh and Lin (2000) analysis of the composite event, the weighted analysis, and a naive analysis (Cook *et al.*, 1996). The estimated correlation matrix between the marginal test statistics based on (3.7) is

$$\begin{bmatrix} 1.000 & 0.412 & 0.276 & 0.008 \\ 0.412 & 1.000 & 0.193 & 0.150 \\ 0.276 & 0.193 & 1.000 & 0.093 \\ 0.008 & 0.150 & 0.093 & 1.000 \end{bmatrix} \tag{5.1}$$

which is instrumental in defining the weights given in the last column of Table 3. The marginal results suggest a trend towards a reduction in the average number of each type of event with pamidronate. The estimated correlations tend to be quite small suggesting that the Simes (1986) procedure may not be excessively conservative in this setting if adjusted marginal analyses are of interest. To test the overall null hypothesis of common marginal mean functions for each type of event ($H_0 : \bigcap_{k=1}^4 H_{k0}$), Simes' (1986) test leads to rejection of H_0 at the first step since $p_{(1)} = 0.001 < \alpha/4 = 0.0125$.

Analyses based on the composite recurrent event give a p-value of 0.062, while an analysis based on the optimally weighted global statistic gives $p = 0.037$. The naive test based on the composite recurrent event suggests the strongest evidence of a difference in the marginal means. Figure 3 displays the naive and adjusted estimates for cumulative mean functions for the composite recurrent event defined by any skeletal complication. As expected, it reveals that the marginal mean functions for the treatment and control groups are over estimated by naive Nelson–Aalen method.

The new global statistic is constructed based on the covariance matrix estimated by (3.7). In principle, it is possible for negative weights to occur. Pocock *et al.* (1987) point out that this happens when the component endpoints are of a quite different nature and there is considerable variability in the pairwise correlations. It may be argued that the use of global test statistics is not ideal in such settings. For

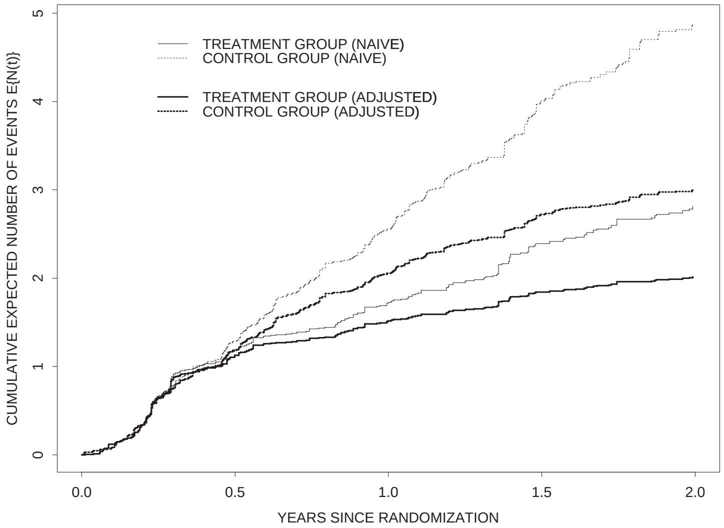

Fig. 3. Naive estimates and survival adjusted estimates for the marginal cumulative mean functions of composite recurrent event from Theriault *et al.* (1999)

the application we consider, the endpoints were chosen to represent adverse events which arise as a consequence of metastatic bone involvement. Since it is expected that more extensive bone involvement will lead to more of each type of skeletal complications, it is not surprising that we did not find negative weights. A graphical display of the nature of this treatment effect on each component skeletal complication is given in Figure 4 where the marginal cumulative mean functions are plotted against time. This plot reveals that, as one would expect based on Table 3, the event type most affected by pamidronate is the need for radiation for bone pain. Use of pamidronate also tends to reduce the marginal cumulative mean number of all other types of events as well.

6. Discussion

We have described a strategy for testing for treatment effects in the context of multivariate recurrent events with a dependent terminal event. The general strategy is to construct marginal test statistics for each type of recurrent event while adjusting for the possibility of dependent termination, and then to synthesize the evidence across all event types by constructing a linear global test statistic. The simulation results revealed empirical type I error rates of the proposed methods which were in close agreement with the nominal levels. This is in contrast to the naive methods which are based on the assumption that the recurrent event process is independently terminated; these naive approaches may underestimate the standard errors of the test statistics and hence increase the risk of a type I error above the nominal level. We also found that there are settings in which the global tests lead to more powerful assessments of treatment effects than analyses based on a composite recurrent event. This occurs when the treatment reduces the incidence of the less frequent events. In contrast, when the treatment effects were manifested on the more frequent event types, analyses based on a composite endpoint lead to more powerful tests. A referee has pointed out that a weighted combination of the test statistic based on a composite event and the global test statistic may provide a means of hypothesis testing which is less sensitive to the variation of the treatment effect across the event types. Exploration of this idea is beyond the scope of this paper but is worthy of future research. When the proportional reduction in the rate of events is the same for each event type, tests based on composite recurrent events and global statistics lead to comparable power.

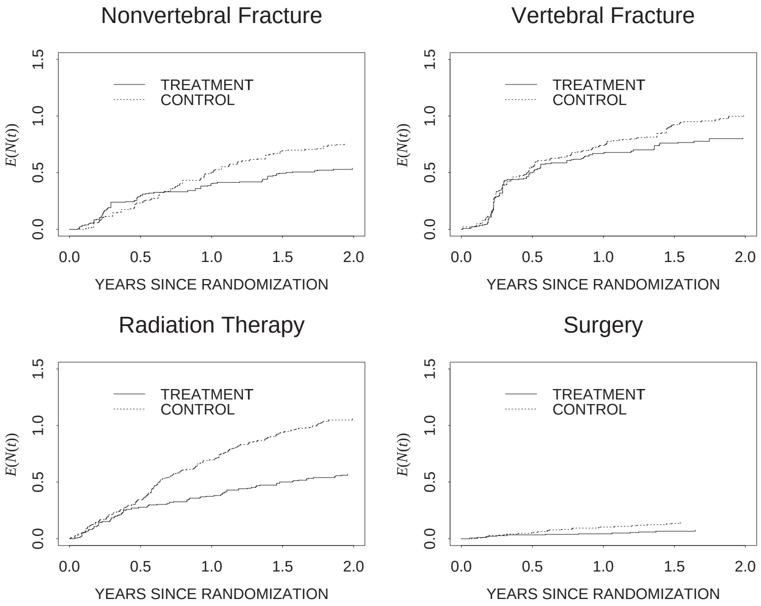

Fig. 4. Marginal cumulative mean functions for nonvertebral fractures, vertebral fractures, need for radiation therapy, and need for surgery for data from Theriault *et al.* (1999)

The methods proposed have the desirable property that marginal analyses are conducted as a by-product of constructing a global test statistic. Unlike approaches based on composite events which only show overall treatment effect, global tests provide information on how the treatment effects each type of event. If a treatment reduces the incidence of one type of event and increases the incidence of another, the nature of these effects may be lost if analyses are based on a composite event, but not when marginal analyses are conducted. Software for implementing these global tests are available from the authors upon request.

Ghosh and Lin (2000) propose the use of global tests which respond to treatment effects on both the marginal event rate and survival. A similar extension would be straightforward to develop here, however, we favor the use of separate tests of treatment effect for the recurrent events and survival since they are directed at two rather different questions.

We have stressed throughout this paper that the marginal tests discussed are directed at detecting treatment effects at the population level, which thereby implies that they are most useful for questions related to overall disease burden to the health care system. Health economists and policy makers are frequently interested in such quantities when the events are associated with health resource utilization and hence costs. Even from this perspective, however, it is essential that assessments of this sort are made in conjunction with careful examination of possible differences in the survival distributions between groups.

Alternative approaches for formulating treatment effects on the recurrent event process may be based on a model for the recurrent events conditional on survival time (Cook and Lawless, 1997). Such conditional models seem unnatural to some degree since they are conditional on a possibly latent variable, and this variable is itself a response to treatment. Models of this sort are used in the context of incomplete longitudinal data however, where they are termed pattern-mixture models. In settings with recurrent events, they generate measures of treatment effect which are easier to relate to individual patients. Specifically, they admit estimates of relative rates for treated versus untreated patients with comparable survival times. Chen and Cook (2004) consider regression-based and stratified models in this spirit which are fit using a modified EM-algorithm.

ACKNOWLEDGEMENTS

This research was supported by the Natural Sciences and Engineering Research Council of Canada and the Canadian Institute for Health Research (CIHR). R. J. Cook is a CIHR Investigator. We thank Dr. John Seaman and Dr. Bee Chen for providing the data from the pamidronate trial, Ms. Ker-Ai Lee for programming assistance, and Dr. Pierre Major for comments and useful discussions. Helpful comments from referees and an Associate Editor are also gratefully acknowledged.

APPENDIX A

We describe here a straightforward generalization of Ghosh and Lin (2000)'s result for multivariate recurrent event process. Define the counting process $D_i(t) = I(X_i \leqslant t, \delta_i = 1)$, let $Y_{\cdot}(s) = \sum_{i=1}^m Y_i(s)$ be the total number of subject at risk for time s, then the Nelson–Aalen estimate of the cumulative hazard function for $D_i(t)$ is given by $\hat{H}(t) = \sum_{i=1}^m \int_0^t Y_{\cdot}(u)\mathrm{d}D_i(u)$. Let $M_{ki}(t) = N_{ki}(t) - \int_0^t Y_i(u)\mathrm{d}\hat{R}_k(u)$ $(k = 1, 2, \ldots, K)$ for type k recurrent event; and $M_i^D(t) = D_i(t) - \hat{H}(t)$, following Ghosh and Lin (2000) we can show that the $\sqrt{m}\{\hat{\mu}_k(t) - \mu_k(t)\} = m^{-\frac{1}{2}} \sum_{i=1}^m \Psi_{ki}(t) + o_p(1)$ with

$$\Psi_{ki}(t) = \int_0^t \frac{S(u)\mathrm{d}M_{ki}(u)}{\pi(u)} - \int_0^t \frac{\{\mu_k(t) - \mu_k(u)\}\mathrm{d}M_i^D(u)}{\pi(u)}, \tag{A.1}$$

where $\pi(u) = \Pr(X \geqslant u)$ and $S(u)$ is survival function of terminal event times. For fixed t, the multivariate central limit theorem implies that $\sqrt{m}\{\hat{\boldsymbol{\mu}}(t) - \boldsymbol{\mu}(t)\}$ follows a mean zero multivariate normal distribution. The asymptotic covariance between jth component and kth component is given by $\sigma_{jk} = E\{\Psi_{j1}(t)\Psi_{k1}(t)\}$, which can be consistently estimated by replacing all the unknown quantities with their corresponding empirical estimates.

Now we will show the structure of asymptotic covariance matrix of $(\hat{Q}_1, \hat{Q}_2, \ldots, \hat{Q}_K)$. Based on Ghosh and Lin (2000)'s results, we know that under the null hypothesis $\mu_{k1}(t) = \mu_{k0}(t)$ for all $0 \leqslant t \leqslant \tau$, $\hat{Q}_k$ can be decomposed to give

$$\begin{aligned}
&\sqrt{\frac{m_1 m_0}{m}} \int_0^\tau \hat{W}_m(t)\mathrm{d}\{\hat{\mu}_{k1}(t) - \hat{\mu}_{k0}(t)\} \\
&= \sqrt{\frac{m_1 m_0}{m}} \left[\int_0^\tau W(t)\mathrm{d}\{\hat{\mu}_{k1}(t) - \mu_{k1}(t)\} - \int_0^\tau W(t)d\{\hat{\mu}_{k0}(t) - \mu_{k0}(t)\}\right] + o_p(1) \\
&= \frac{1}{\sqrt{m}} \sum_{\ell=0}^{1} (-1)^{(1-\ell)} \sqrt{\frac{m_{1-\ell}}{m_\ell}} \left[\sum_{i=1}^m \left\{\int_0^\tau W(t)\mathrm{d}\Psi_{ki}(t)\right\} I(z_i = \ell)\right] + o_p(1),
\end{aligned}$$

where z_i is a binary covariate such that $z_i = 1$ for subjects in the treated group and $z_i = 0$ for subjects in the control group. It is clear that the above expression is a sum of independent identically distributed random variables with mean zero. Thus, the central limit theorem ensures that $(\hat{Q}_1, \hat{Q}_2, \ldots, \hat{Q}_K)$ has a multivariate normal asymptotic distribution with mean zero and covariance between $\hat{Q}_j$ and $\hat{Q}_k$ is given by

$$\operatorname{cov}(\hat{Q}_j, \hat{Q}_k) = \sum_{\ell=0}^{1} \sqrt{\rho_{1-\ell}}\, E\left\{\int_0^\tau W(t)\mathrm{d}\Psi_{ji}(t) \int_0^\tau W(t)\mathrm{d}\Psi_{ki}(t) I(z_i = \ell)\right\}, \tag{A.2}$$

for $1 \leqslant j, k \leqslant K$, which can be consistently estimated by

$$\frac{1}{m}\sum_{\ell=0}^{1}\frac{m_{1-\ell}}{m_\ell}\sum_{i=1}^{m}\left\{\int_0^\tau \hat{W}(t)\mathrm{d}\hat{\Psi}_{ji}(t)\int_0^\tau \hat{W}(t)\mathrm{d}\hat{\Psi}_{ki}(t)\right\}I(z_i=\ell). \tag{A.3}$$

Ghosh and Lin (2000) provide details on how to decompose $\{\hat{R}_{jk}(u) - R_{jk}(u)\}$ and $\{\hat{S}(u) - S(u)\}$.

References

ARMITAGE, P. AND PARMAR, M. (1986). Some approaches to the problem of multiplicity in clinical trials. *Proceedings of the XIIth International Biometrics Conference,* Seattle: Biometric Society.

BANG, H. AND TSIATIS, A. A. (2000). Estimating medical costs with censored data. *Biometrika* **87**, 329–343.

CHEN, E. B. AND COOK, R. J. (2004). Stratified analysis of event history data with dependent censoring. *Technical Report*. University of Waterloo.

COOK, R. J., LAWLESS, J. F. AND NADEAU, C. (1996). Robust tests for treatment comparisons based on recurrent responses. *Biometrics* **52**, 557–571.

COOK, R. J. AND LAWLESS, J. F. (1997). Marginal analysis of recurrent events and a terminating event. *Statistics in Medicine* **16**, 911–924.

GHOSH, D. AND LIN, D. Y. (2000). Nonparametric analysis of recurrent events and death. *Biometrics* **56**, 554–562.

JAMES, S. (1991). Approximate multinormal probabilities applied to correlated multiple endpoints in clinical trials. *Statistics in Medicine* **10**, 1123–1135.

LI, Q. H. AND LAGAKOS, S. W. (1997). Use of the Wei–Lin–Weissfeld method for the analysis of a recurring and a terminating event. *Statistics in Medicine* **16**, 925–940.

LIN, D. Y., FEUER, E. J., ETZIONI, R. AND WAX, Y. (1997). Estimating medical costs from incomplete follow-up data. *Biometrics* **53**, 419–434.

O'BRIEN, P. C. (1984). Procedures for comparing samples with multiple endpoints. *Biometrics* **40**, 1079–1087.

POCOCK, S. J., GELLER, N. L. AND TSIATIS, A. A. (1987). The analysis of multiple endpoints in clinical trials. *Biometrics* **43**, 487–498.

RAO, C. R. (1973). *Linear Statistical Inference and its Applications,* 2nd edn. New York: Wiley.

SIMES, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika* **73**, 751–754.

STRAWDERMAN, R. (2000). Estimating the mean of an increasing stochastic processes at a censored stopping time. *Journal of the American Statistical Association* **95**, 1192–1298.

THERIAULT, R. L., LIPTON, A., HORTOBAGYI, G. N., LEFF, R., GLÜCK, S., STEWARD, J. F., COSTELLO, S., KENNEDY, I. SIMEONE, J. *et al.* for the Protocol 18 Aredia Breast Cancer Study Group. Pamidronate reduces skeletal morbidity in women with advanced breast cancer and lytic bone lesions: A randomized, placebo-controlled trial (1999). *Journal of Clinical Oncology* **17**, 846–854.

WEI, L. J., LIN, D. Y. AND WEISSFELD, L. (1989). Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *Journal of the American Statistical Association* **84**, 1065–1073.

ZHAO, H. AND TSIATIS, A. A. (1997). A consistent estimator for the distribution of quality adjusted survival time. *Biometrika* **84**, 339–348.

ZHAO, H. AND TSIATIS, A. A. (1999). Efficient estimation of the distribution of quality-adjusted survival time. *Biometrics* **55**, 1101–1107.

[*Received February 19, 2002; first revision March 3, 2003; second revision August 27, 2003;
accepted for publication September 2, 2003*]